TIME SERIES FORECASTING –SPARKLING WINE

DSBA

SUBMITTED BY SAUMYA JAIN (PGP-DSBA)

Contents

List of Tables:

List of Plots:

Problem:

For this assignment, the data of different types of wine sales in the 20th century is to be analysed. Both data are from the same company but of different wines. As an analyst in the ABC Estate Wines, you are tasked to analyse and forecast Wine Sales in the 20th century.

Data set for the Problem: [Sparkling.csv]

1. Read the data as an appropriate Time Series data and plot the data.

- The dataset given to us that contains the information of sales of rose wine. The excel has 187 rows and 1 column.
- We have also set index to be Year Month.
- The description of the dataset is as below.

	count	mean	std	min	25%	50%	75%	max
Sparkling	187.0	2402.417112	1295.11154	1070.0	1605.0	1874.0	2549.0	7242.0

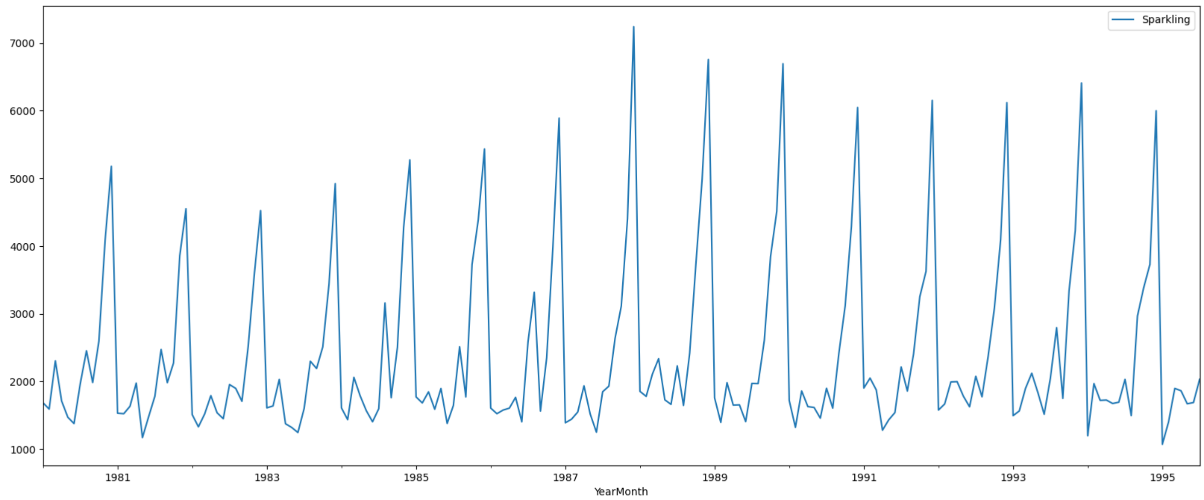
- We have seen the top 5 rows of the dataset shared as shown below:

YearMonth	Sparkling
1980-01-01	1686
1980-02-01	1591
1980-03-01	2304
1980-04-01	1712
1980-05-01	1471

- We have seen the last 5 rows of the dataset shared as shown below:

	Sparkling
YearMonth	
1995-03-01	1897
1995-04-01	1862
1995-05-01	1670
1995-06-01	1688
1995-07-01	2031

- Plot the graph:



2. Perform appropriate Exploratory Data Analysis to understand the data and perform decomposition.

- Datatypes of the data present in Rose wine dataset.

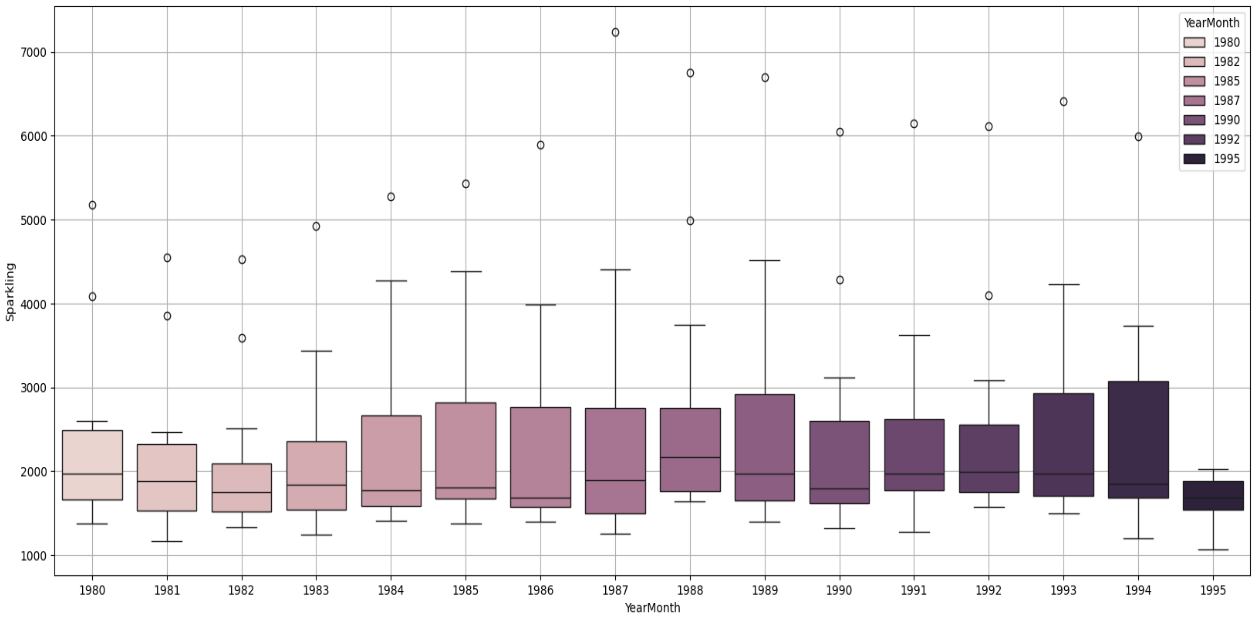
```
Sparkling    int64
dtype: object
```

- Check for null values present in the given dataset and found that there are no null values present in dataset as shown below:

```
Sparkling    0
dtype: int64
```

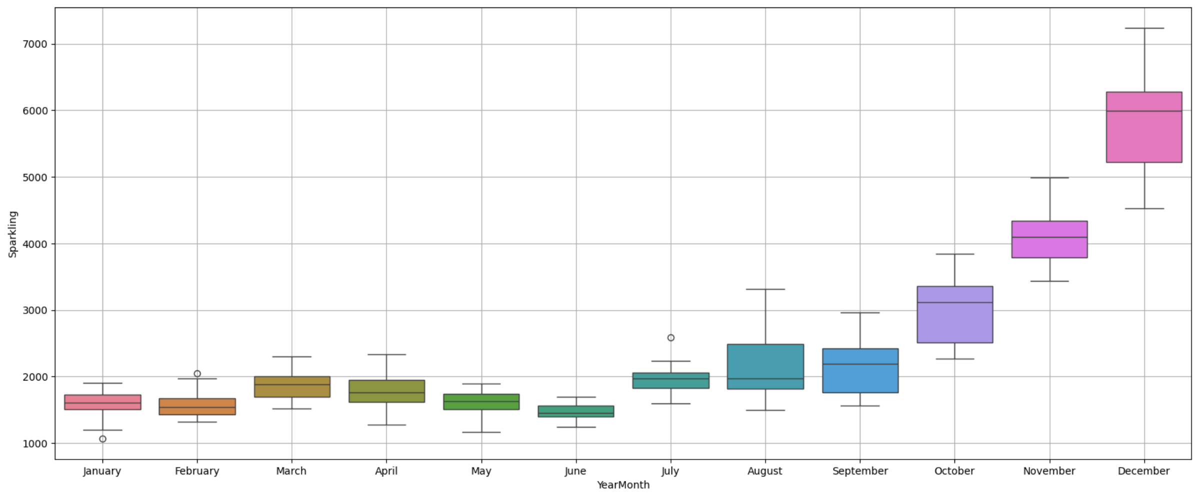
- **Boxplot Yearly:**

The boxplot yearly shows that the there is peak in 1988-1989 . Though outliers are also present in all years. However the boxplot shows that there is consistency in all years.



- **Monthly Boxplot**

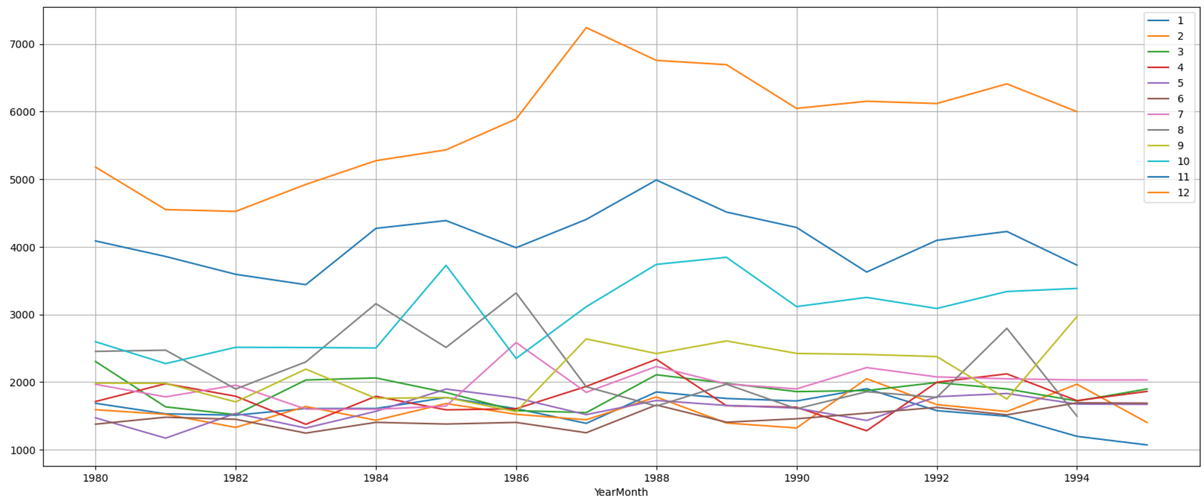
1. From the graph we can inference that the sales of sparkling wine is mostly high in December and lowest in January. Outliers are present in Jan ,February, July.

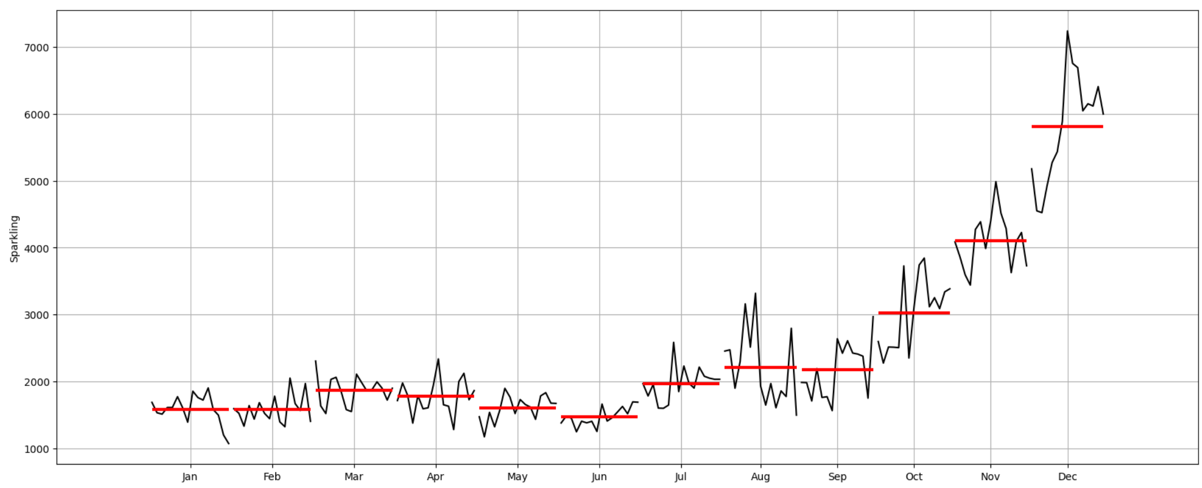


- **Graph of Monthly Sales over the years:**

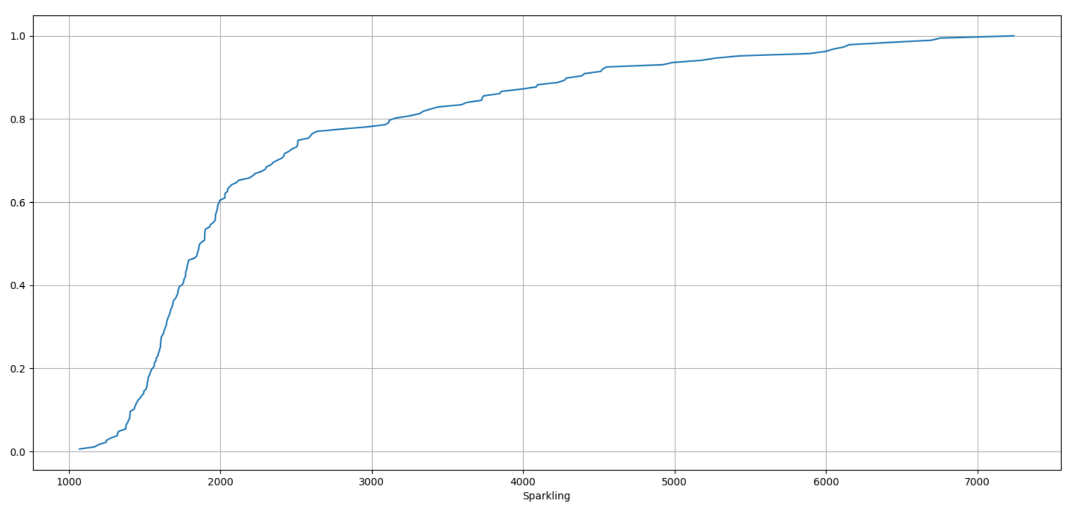
The sales of Sparkling wine is highest in 12th Month i.e. December and the year 1988 was the year with the highest number of sales.

YearMonth	1	2	3	4	5	6	7	8	9	10	11	12
YearMonth												
1980	1686.0	1591.0	2304.0	1712.0	1471.0	1377.0	1966.0	2453.0	1984.0	2596.0	4087.0	5179.0
1981	1530.0	1523.0	1633.0	1976.0	1170.0	1480.0	1781.0	2472.0	1981.0	2273.0	3857.0	4551.0
1982	1510.0	1329.0	1518.0	1790.0	1537.0	1449.0	1954.0	1897.0	1706.0	2514.0	3593.0	4524.0
1983	1609.0	1638.0	2030.0	1375.0	1320.0	1245.0	1600.0	2298.0	2191.0	2511.0	3440.0	4923.0
1984	1609.0	1435.0	2061.0	1789.0	1567.0	1404.0	1597.0	3159.0	1759.0	2504.0	4273.0	5274.0
1985	1771.0	1682.0	1846.0	1589.0	1896.0	1379.0	1645.0	2512.0	1771.0	3727.0	4388.0	5434.0
1986	1606.0	1523.0	1577.0	1605.0	1765.0	1403.0	2584.0	3318.0	1562.0	2349.0	3987.0	5891.0
1987	1389.0	1442.0	1548.0	1935.0	1518.0	1250.0	1847.0	1930.0	2638.0	3114.0	4405.0	7242.0
1988	1853.0	1779.0	2108.0	2336.0	1728.0	1661.0	2230.0	1645.0	2421.0	3740.0	4988.0	6757.0
1989	1757.0	1394.0	1982.0	1650.0	1654.0	1406.0	1971.0	1968.0	2608.0	3845.0	4514.0	6694.0
1990	1720.0	1321.0	1859.0	1628.0	1615.0	1457.0	1899.0	1605.0	2424.0	3116.0	4286.0	6047.0
1991	1902.0	2049.0	1874.0	1279.0	1432.0	1540.0	2214.0	1857.0	2408.0	3252.0	3627.0	6153.0
1992	1577.0	1667.0	1993.0	1997.0	1783.0	1625.0	2076.0	1773.0	2377.0	3088.0	4096.0	6119.0
1993	1494.0	1564.0	1898.0	2121.0	1831.0	1515.0	2048.0	2795.0	1749.0	3339.0	4227.0	6410.0
1994	1197.0	1968.0	1720.0	1725.0	1674.0	1693.0	2031.0	1495.0	2968.0	3385.0	3729.0	5999.0
1995	1070.0	1402.0	1897.0	1862.0	1670.0	1688.0	2031.0	NaN	NaN	NaN	NaN	NaN





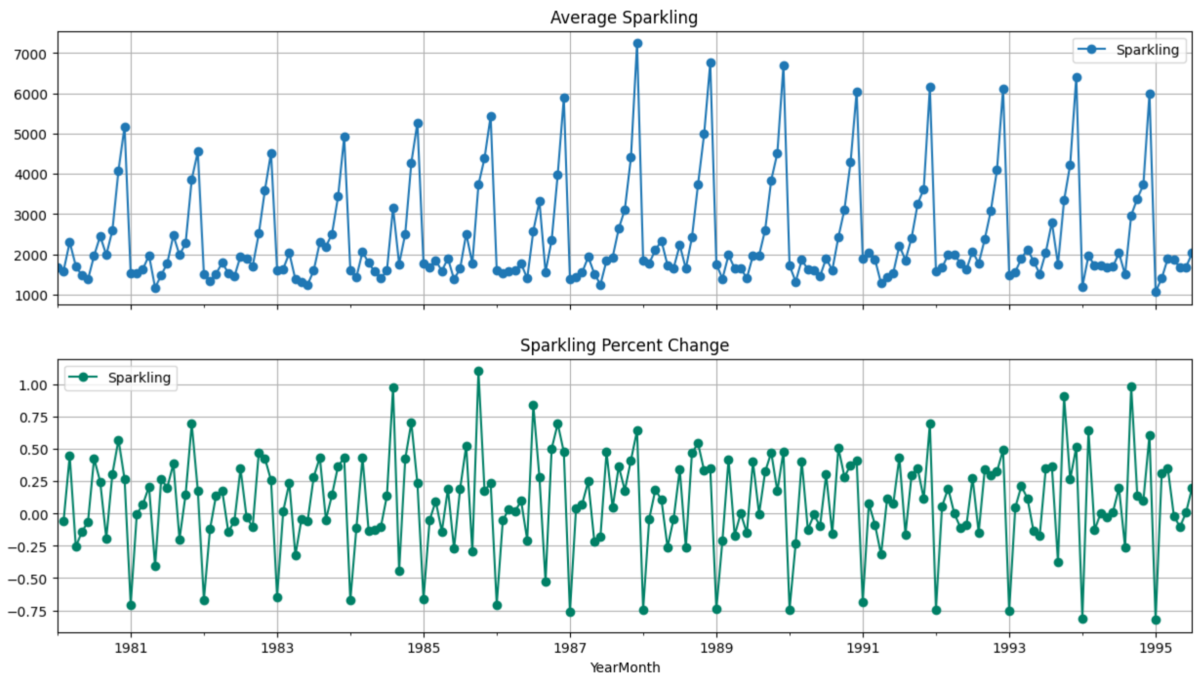
- Plot the Empirical Cumulative Distribution.



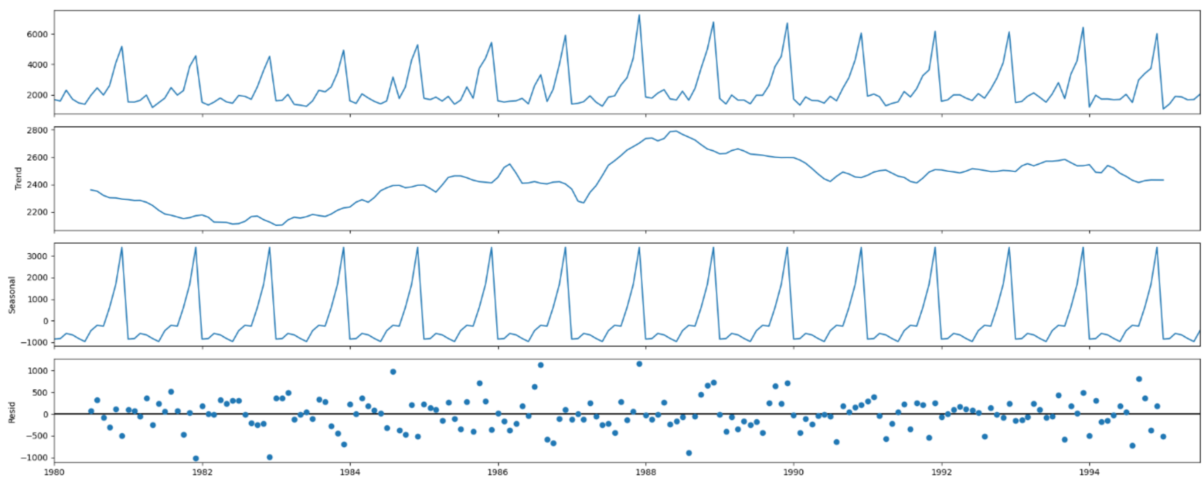
The interference from Empirical Cumulative Distribution is:

- Highest value is 7000.
- Approximately 80% of sales is less than 3000.
- More than 50% of sales have been less than 2000.

- ## **Average Sparkling Sales and Average Rose percentage**



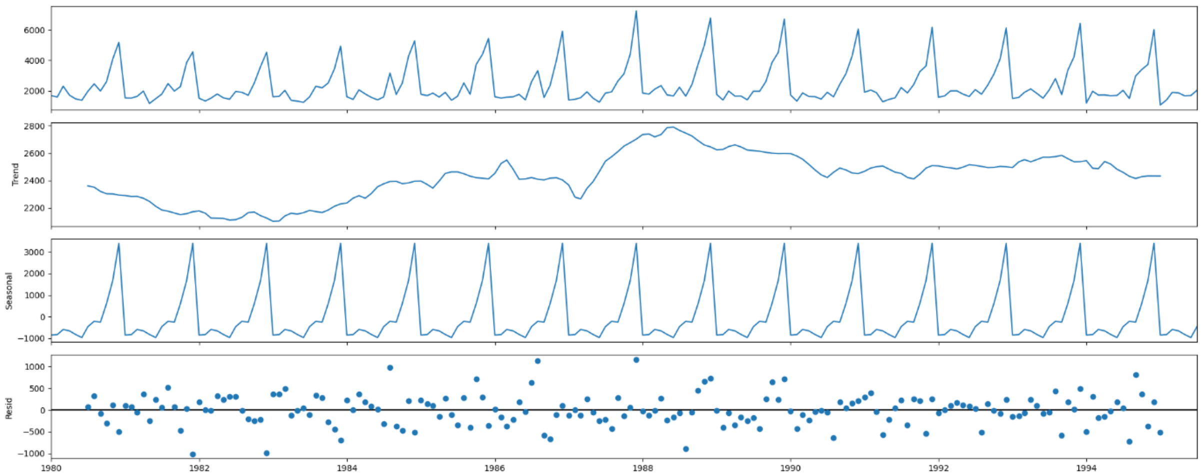
- ## **Decomposition --- Additive:**



Decomposition by additive plot shows that:

- It shows that the trend has been declined after 1988-1989.
- Peak year was 1988-1989.
- Residue is spread not showing a pattern
- Seasonality and trend is present.

- ## Decomposition --- Multiplicative:



- Decomposition by multiplicative plot shows that:
  - It shows that the trend has been declined after 1988-1989.
  - Peak year was 1988-1989.
  - Residue is between 0 and 1 but in additive it is between 0 and 1000 that is high.
  - Seasonality and trend is present.
  - Multiplicative decomposition is better in this case as the residue is between the range 0 and 1.

- ## Trend, Seasonality and Residue

```
Residual
 YearMonth
1980-01-01           NaN
1980-02-01           NaN
1980-03-01           NaN
1980-04-01           NaN
1980-05-01           NaN
1980-06-01           NaN
1980-07-01     70.835599
1980-08-01    315.999487
1980-09-01    -81.864401
1980-10-01   -307.353290
1980-11-01    109.891154
1980-12-01   -501.775513
Name: resid, dtype: float64
```

-

```
Trend
 YearMonth
1980-01-01            NaN
1980-02-01            NaN
1980-03-01            NaN
1980-04-01            NaN
1980-05-01            NaN
1980-06-01            NaN
1980-07-01    2360.666667
1980-08-01    2351.333333
1980-09-01    2320.541667
1980-10-01    2303.583333
1980-11-01    2302.041667
1980-12-01    2293.791667
Name: trend, dtype: float64

Seasonality
 YearMonth
1980-01-01    -854.260599
1980-02-01    -830.350678
1980-03-01    -592.356630
1980-04-01    -658.490559
1980-05-01    -824.416154
1980-06-01    -967.434011
1980-07-01    -465.502265
1980-08-01    -214.332821
1980-09-01    -254.677265
1980-10-01     599.769957
1980-11-01    1675.067179
1980-12-01    3386.983846
Name: seasonal, dtype: float64
```

3. Split the data into training and test. The test data should start in 1991.

The dataset is being split into training and test dataset. The test data set starts from 1991.

The train dataset has 132 rows and 1 column.

The test dataset has 55 rows and 1 column.

- **Dataset of train as shown below**:

Training Data

YearMonth	Sparkling
1980-01-01	1686
1980-02-01	1591
1980-03-01	2304
1980-04-01	1712
1980-05-01	1471
...	...
1990-08-01	1605
1990-09-01	2424
1990-10-01	3116
1990-11-01	4286
1990-12-01	6047

Test dataset:

```
Test Data
```

YearMonth	Sparkling
1991-01-01	1902
1991-02-01	2049
1991-03-01	1874
1991-04-01	1279
1991-05-01	1432
1991-06-01	1540
1991-07-01	2214
1991-08-01	1857
1991-09-01	2408
1991-10-01	3252
1991-11-01	3627
1991-12-01	6153
1992-01-01	1577
1992-02-01	1667
1992-03-01	1993

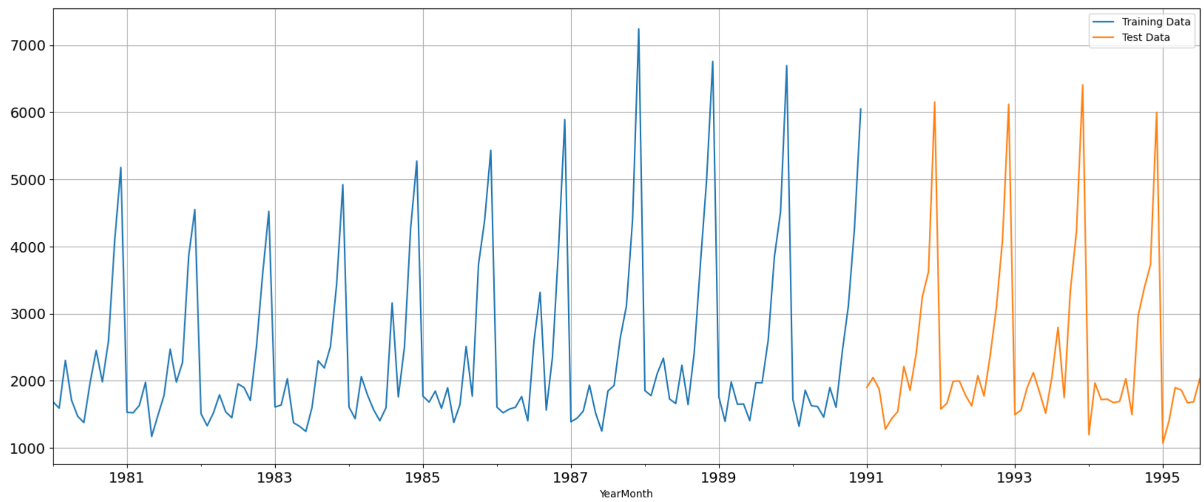
Train datasets describe:

	Sparkling
count	132.000000
mean	2403.780303
std	1303.430250
min	1170.000000
25%	1595.500000
50%	1850.000000
75%	2531.500000
max	7242.000000

Test dataset describe:

	Sparkling
count	55.000000
mean	2399.145455
std	1286.825457
min	1070.000000
25%	1672.000000
50%	1898.000000
75%	2601.500000
max	6410.000000

Plot for training and test dataset:



4. Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression, naive forecast models and simple average models. should also be built on the training data and check the performance on the test data using RMSE.

The below models are built on Training and test dataset.

- Linear Regression
- Naive Approach
- Simple Average
- Moving Average (MA)
- Simple Exponential Smoothing
- Double Exponential Smoothing (Holt's Model)
- Triple Exponential Smoothing (Holt - Winter's Model)

LINEAR REGRESSION:

- Few rows of training and test dataset:

First few rows of Training Data

YearMonth	Sparkling	time
1980-01-01	1686	1
1980-02-01	1591	2
1980-03-01	2304	3
1980-04-01	1712	4
1980-05-01	1471	5

Last few rows of Training Data

YearMonth	Sparkling	time
1990-08-01	1605	128
1990-09-01	2424	129
1990-10-01	3116	130
1990-11-01	4286	131
1990-12-01	6047	132

First few rows of Test Data

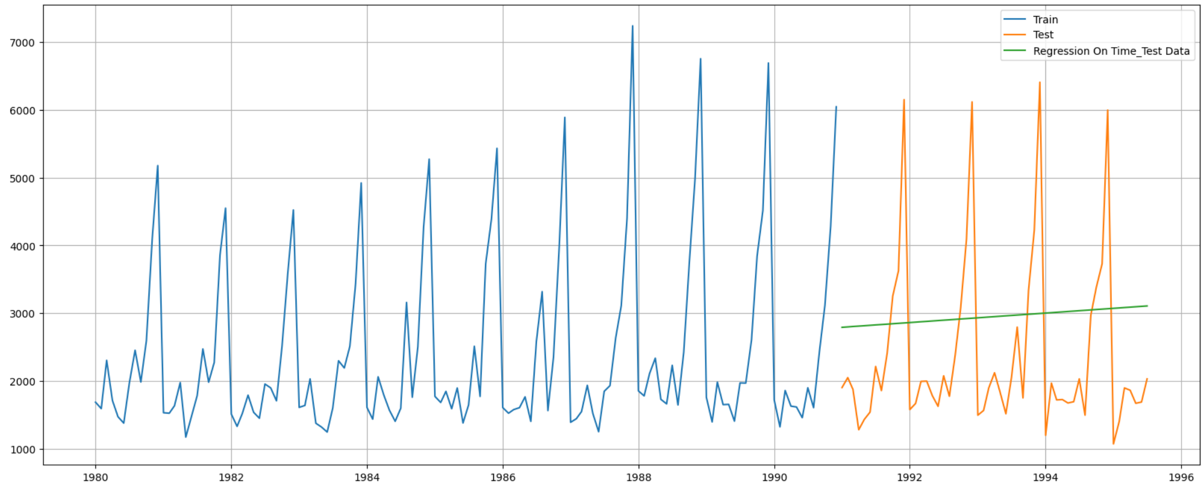
YearMonth	Sparkling	time
1991-01-01	1902	133
1991-02-01	2049	134
1991-03-01	1874	135
1991-04-01	1279	136
1991-05-01	1432	137

Last few rows of Test Data

Sparkling time

Last few rows of Test Data

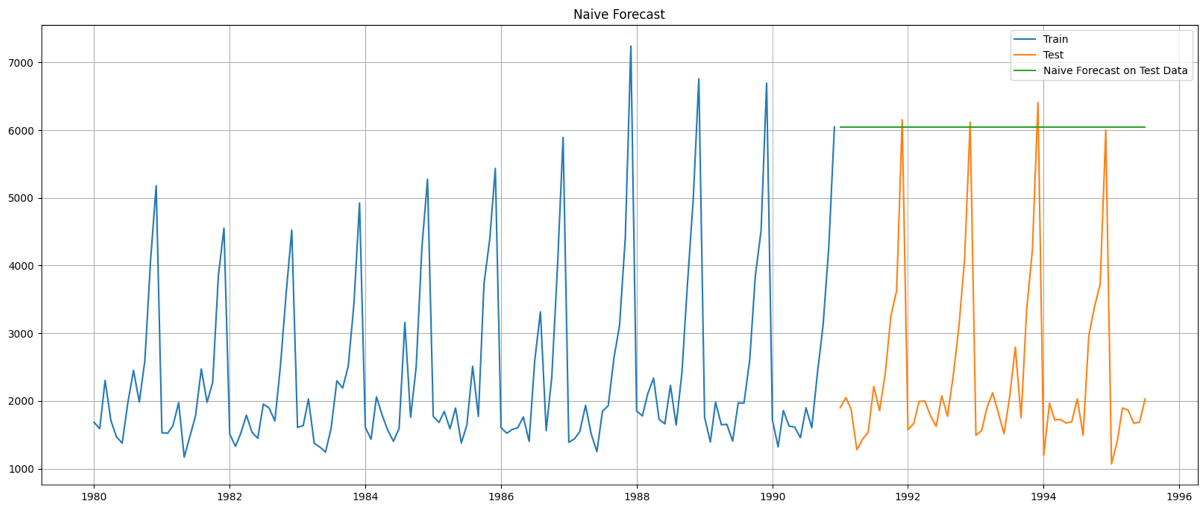
YearMonth	Sparkling	time
1995-03-01	1897	183
1995-04-01	1862	184
1995-05-01	1670	185
1995-06-01	1688	186
1995-07-01	2031	187



- The green line is the prediction made by Linear Model as we can see that the prediction made by the linear model is not good as shown in graph also as the predicted values is far away from the actual values.
- RMSE values calculated for the model is 1389.13

	RMSE
RegressionOnTime	1389.135175

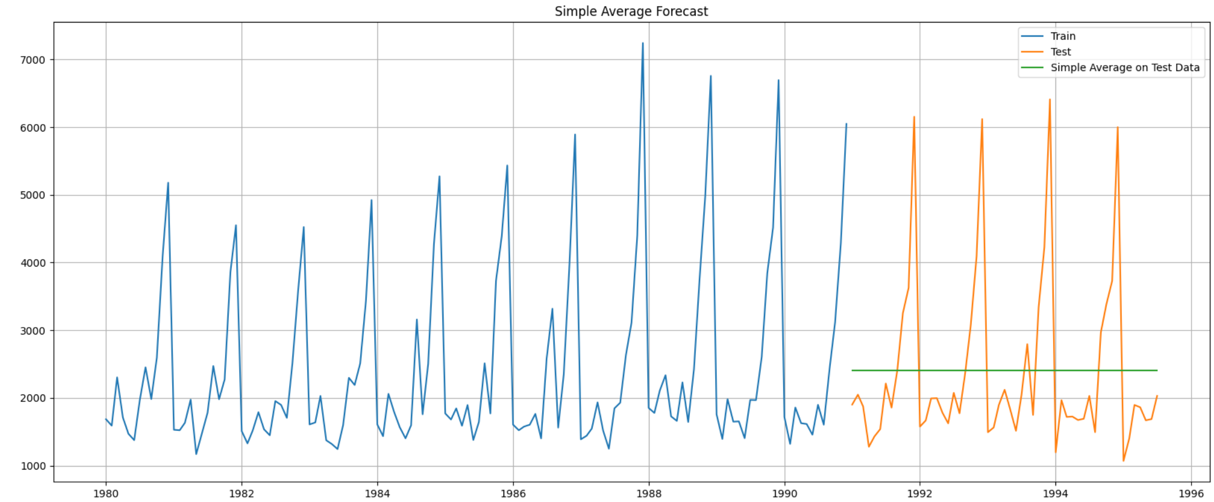
NAIVE'S MODEL:



- The green line is the prediction made by Naive's Model as we can see that the prediction made by the linear model is not good as shown in graph also as the predicted values is far away from the actual values.
- RMSE values calculated for the model is 3864.27. The less the RMSE the better the model.

NaiveModel	3864.279352

- **SIMPLE AVERAGE FORECAST MODEL:**

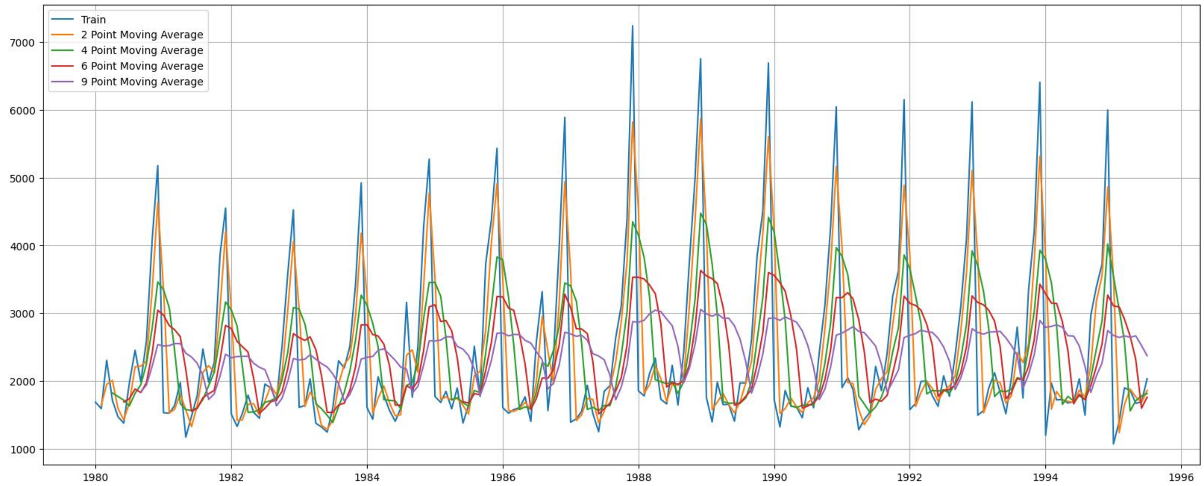


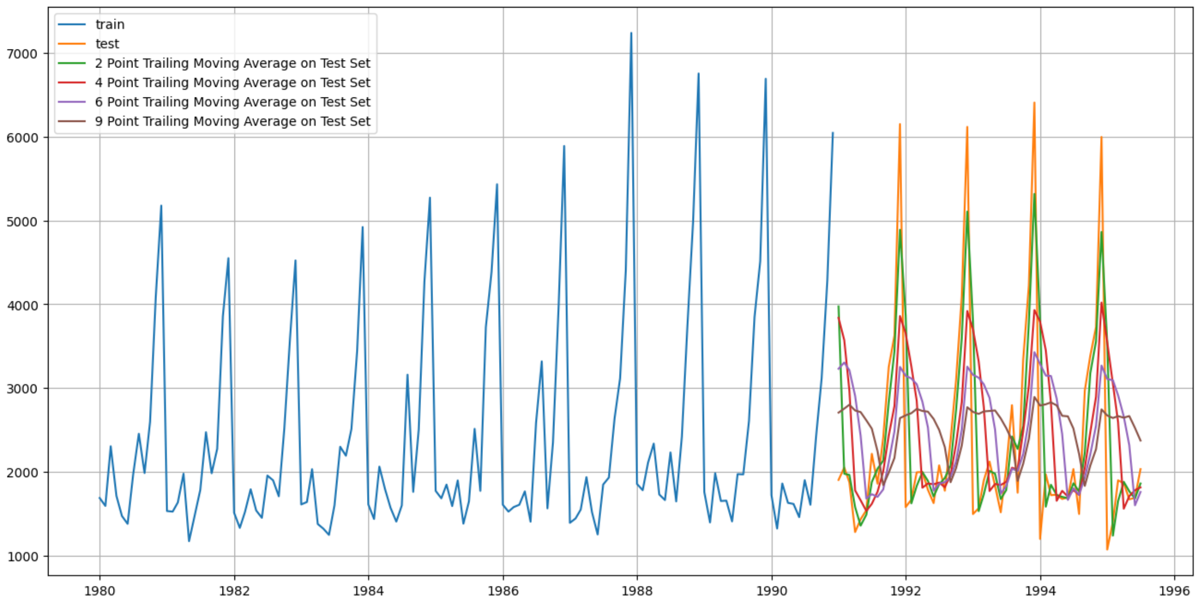
- The green line is the prediction made by Simple Average Forecast Model as we can see that the prediction made by the linear model is not good as shown in graph also as the predicted values is far away from the actual values.
- RMSE values calculated for the model is 1275.08 The less the RMSE the better the model.

SimpleAverageModel	1275.081804

- **MOVING AVERAGE:**

	Sparkling	Trailing_2	Trailing_4	Trailing_6	Trailing_9
YearMonth					
1980-01-01	1686	NaN	NaN	NaN	NaN
1980-02-01	1591	1638.5	NaN	NaN	NaN
1980-03-01	2304	1947.5	NaN	NaN	NaN
1980-04-01	1712	2008.0	1823.25	NaN	NaN
1980-05-01	1471	1591.5	1769.50	NaN	NaN





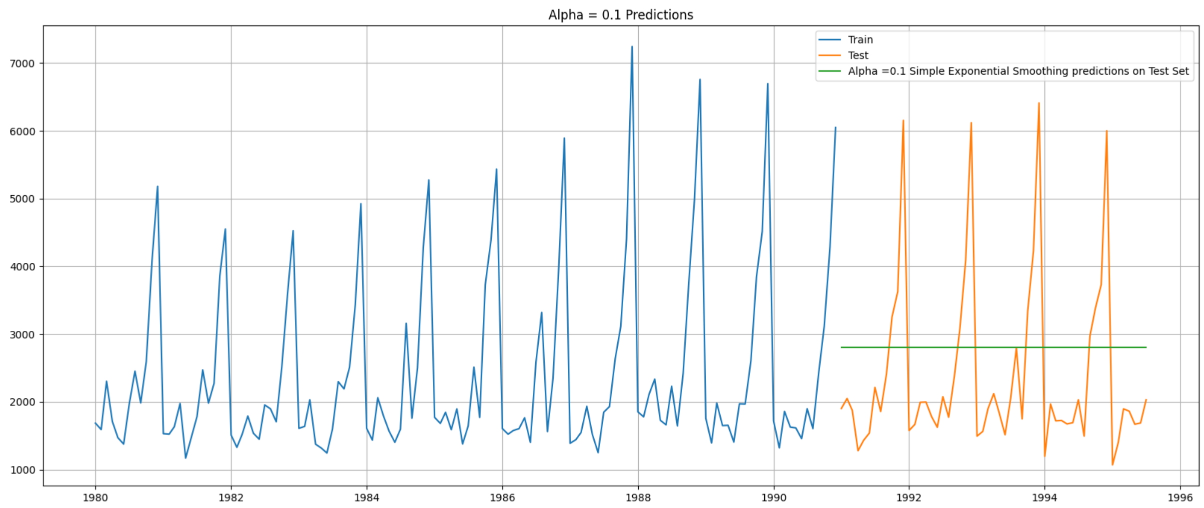
- A moving average model is used for forecasting future values, while moving average smoothing is used for estimating the trend-cycle of past values. Higher the rolling window, smoother will be its curve more values are being taken into account.

- RMSE values calculated for the model are as below. The less the RMSE the better the model.

	RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
6pointTrailingMovingAverage	1283.927428
9pointTrailingMovingAverage	1346.278315

- **SIMPLE EXPONENTIAL:**

Taken all values from 0.1 to 0.9 to find the best alpha value for SIMPLE EXPONENTIAL which has less RMSE .



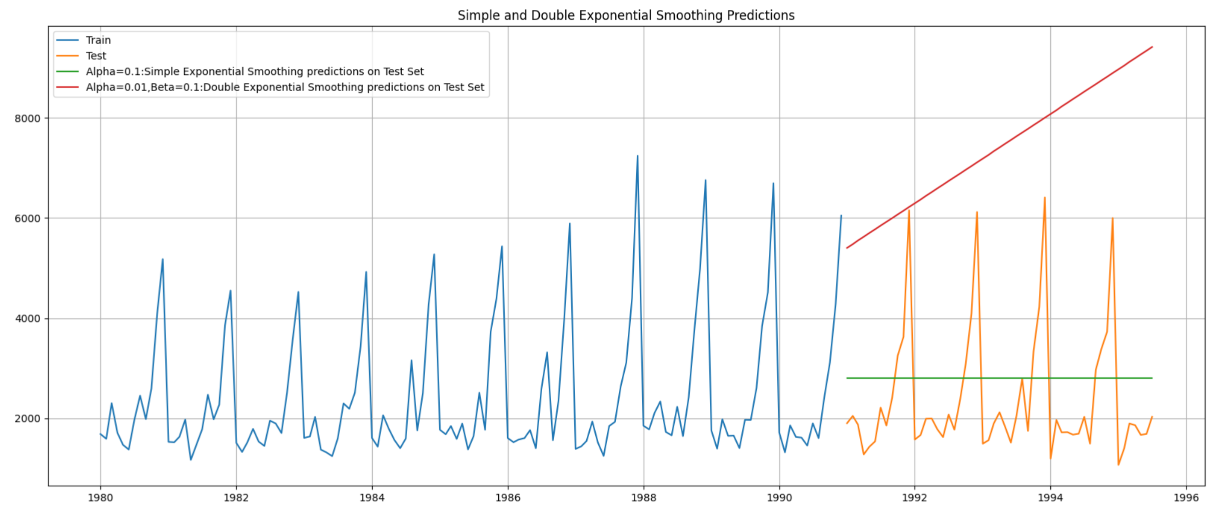
RMSE:

The alpha value 0.1 is giving us less RMSE that is 1338.00 in all the apha values .

Alpha=0.1,SES	1338.004623

- **Double Exponential Smoothing (Holt's Model) :**

Taken all values from 0.1 to 0.9 to find the best alpha , beta value for DOUBLE EXPONENTIAL which has less RMSE .



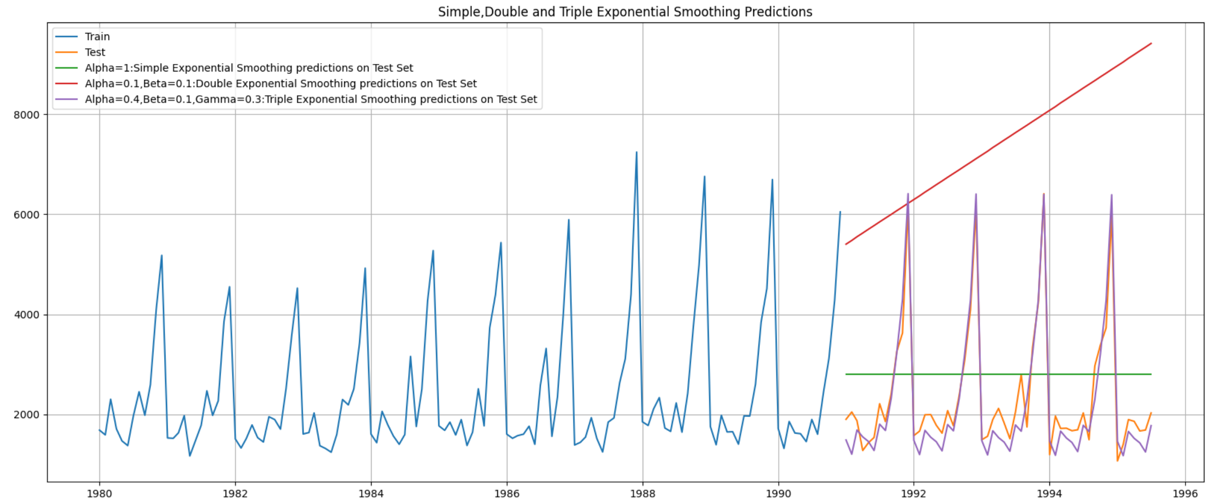
RMSE:

The alpha value and beta 0.1 is giving us less RMSE that is 5291.87 in all the apha , beta values . So best value for alpha, beta is 0.1.

Alpha=0.1,Beta=0.1:DES	5291.879833

- **Triple Exponential Smoothing (Holt Winter's Model) :**

Taken all values from 0.1 to 0.9 to find the best alpha , beta , gamma value for triple exponential smoothing to see which has less RMSE . We can see that the predicted value that is green is fitting the actual values much better than the other models



- **RMSE:**

The alpha value 0.4, beta 0.1 and gamma 0.3 is giving us less RMSE that is 378.62 in all the apha , beta and gamma values . So best value for alpha, beta , gamma is that only.

Alpha=0.4,Beta=0.1,Gamma=0.3:TES 378.626241

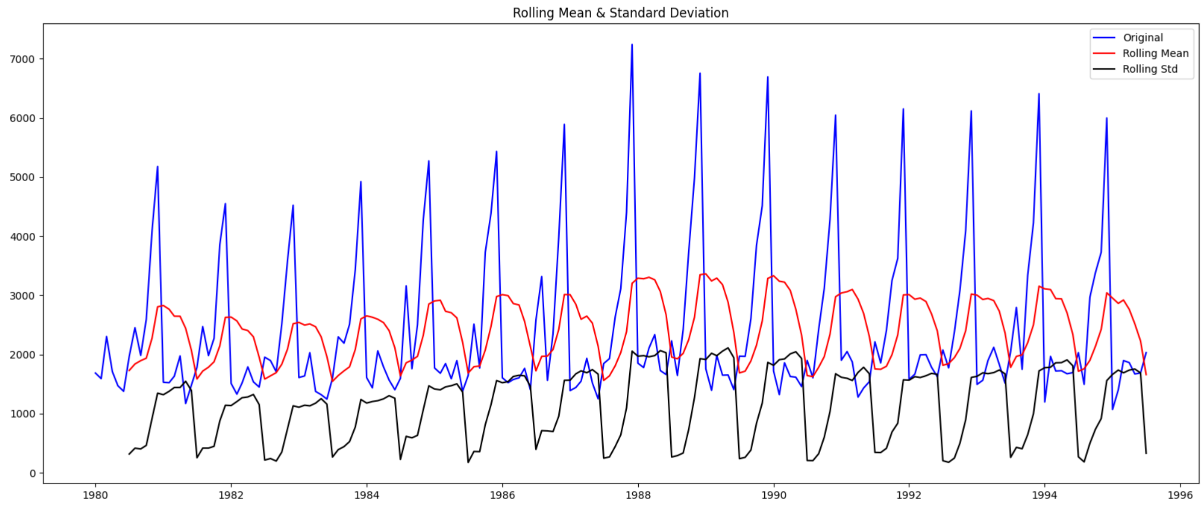
5. Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment. Note: Stationarity should be checked at alpha = 0.05.

For checking the series is stationary or not we have to use Augumented Dickey – Fuller test for the same.

The hypothesis for this is :

If the value is less than 0.05 then the series is stationary and good to move for further ARIMA/SARIMA Model.

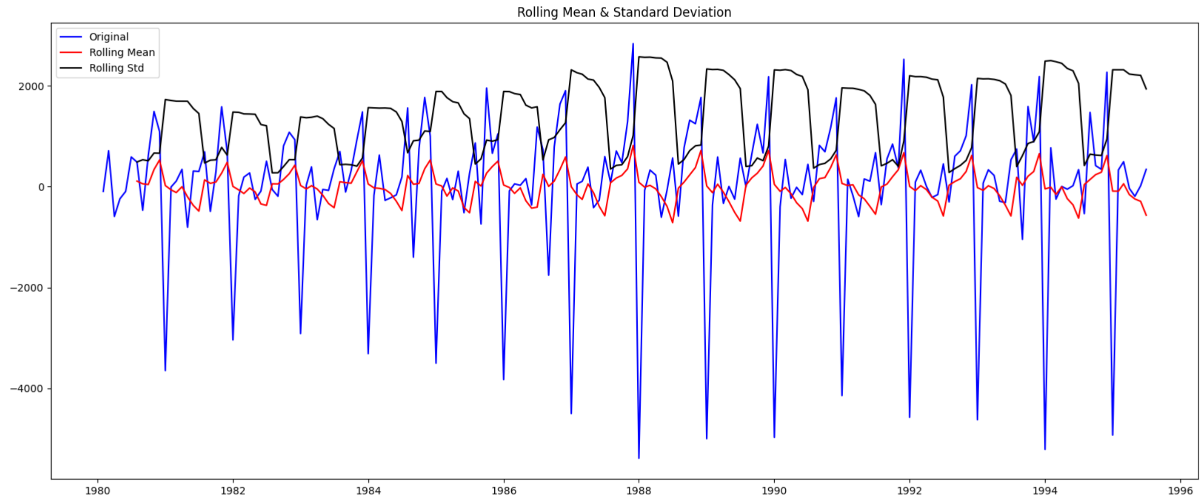
If the value of p value is more than 0.05 then we fail to reject the null hypothesis and the series is not stationary and can't proceed with ARIMA/SARIMA model.




```
Results of Dickey-Fuller Test:
Test Statistic                  -1.360497
p-value                          0.601061
#Lags Used                      11.000000
Number of Observations Used    175.000000
Critical Value (1%)             -3.468280
Critical Value (5%)             -2.878202
Critical Value (10%)            -2.575653
dtype: float64
```

Let us take a difference of order 1 and check whether the Time Series is stationary or not.

We used .diff() function on the existing series without any argument, implying the default diff value of 1 and also dropped the NaN values, since differencing of order 1 would generate the first value as NaN which need to be dropped



We can see that now the p value is 0 that is much smaller than the 0.05 so we fail to reject the null hypothesis and considering the series as stationary and good to move further for ARIMA / SARIMA Model as the series is stationary.

```
Results of Dickey-Fuller Test:
Test Statistic                 -45.050301
p-value                          0.000000
#Lags Used                      10.000000
Number of Observations Used    175.000000
Critical Value (1%)             -3.468280
Critical Value (5%)             -2.878202
Critical Value (10%)            -2.575653
dtype: float64
```

6. Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE.

The values of p,q,d where p is the order of AR , q is the order of Moving average and d is the difference that will make the series stationary for this a for loop has been there .

```
Some parameter combinations for the Model...
Model: (0, 1, 1)
Model: (0, 1, 2)
Model: (1, 1, 0)
Model: (1, 1, 1)
Model: (1, 1, 2)
Model: (2, 1, 0)
Model: (2, 1, 1)
Model: (2, 1, 2)
```

Less the AIC we will take that model in this case 2,1,2 has the lowest AIC so we need to sort the AIC.

```
ARIMA(0, 1, 0) - AIC:2267.6630357855465
ARIMA(0, 1, 1) - AIC:2263.0600155919765
ARIMA(0, 1, 2) - AIC:2234.4083231352784
ARIMA(1, 1, 0) - AIC:2266.6085393190097
ARIMA(1, 1, 1) - AIC:2235.755094674255
ARIMA(1, 1, 2) - AIC:2234.5272004518056
ARIMA(2, 1, 0) - AIC:2260.36574396809
ARIMA(2, 1, 1) - AIC:2233.777626238336
ARIMA(2, 1, 2) - AIC:2213.5092125741553
```

After the sort we found that Less the AIC we will take that model in this case 2,1,3 has the lowest AIC.

	param	AIC
8	(2, 1, 2)	2213.509213
7	(2, 1, 1)	2233.777626
2	(0, 1, 2)	2234.408323
5	(1, 1, 2)	2234.527200
4	(1, 1, 1)	2235.755095
6	(2, 1, 0)	2260.365744
1	(0, 1, 1)	2263.060016
3	(1, 1, 0)	2266.608539
0	(0, 1, 0)	2267.663036

The summary report for the ARIMA Model with values (2,1,2) as p,q,d respectively.

```
                               SARIMAX Results
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  132
Model:                 ARIMA(2, 0, 1)   Log Likelihood               -1113.295
Date:                Fri, 23 Feb 2024   AIC                           2236.591
Time:                        19:28:55   BIC                           2251.005
Sample:                    01-01-1980   HQIC                          2242.448
                         - 12-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const       2399.4586    118.215     20.297      0.000    2167.762    2631.155
ar.L1          1.2375      0.138      8.938      0.000       0.966       1.509
ar.L2         -0.5293      0.124     -4.266      0.000      -0.772      -0.286
ma.L1         -0.8080      0.156     -5.174      0.000      -1.114      -0.502
sigma2      1.233e+06   1.37e+05      9.016      0.000    9.65e+05     1.5e+06
===================================================================================
Ljung-Box (L1) (Q):                   0.03   Jarque-Bera (JB):                26.42
Prob(Q):                              0.86   Prob(JB):                         0.00
Heteroskedasticity (H):               2.40   Skew:                             0.80
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.49
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

- **RMSE for AUTO_ARIMA:**

ARIMA(2,0,1)	1269.345658

- # SARIMA MODEL

SARIMA utilizes a variety of auto-regression (AR) and moving average (MA) models, as well as differencing, to capture trends and seasonality in data.

```
Examples of some parameter combinations for Model...
Model: (0, 1, 1)(0, 0, 1, 12)
Model: (0, 1, 2)(0, 0, 2, 12)
Model: (0, 1, 3)(0, 0, 3, 12)
Model: (1, 1, 0)(1, 0, 0, 12)
Model: (1, 1, 1)(1, 0, 1, 12)
Model: (1, 1, 2)(1, 0, 2, 12)
Model: (1, 1, 3)(1, 0, 3, 12)
Model: (2, 1, 0)(2, 0, 0, 12)
Model: (2, 1, 1)(2, 0, 1, 12)
Model: (2, 1, 2)(2, 0, 2, 12)
Model: (2, 1, 3)(2, 0, 3, 12)
Model: (3, 1, 0)(3, 0, 0, 12)
Model: (3, 1, 1)(3, 0, 1, 12)
Model: (3, 1, 2)(3, 0, 2, 12)
Model: (3, 1, 3)(3, 0, 3, 12)

SARIMA(0, 1, 0)x(0, 0, 0, 12) - AIC:2251.3597196862966
SARIMA(0, 1, 0)x(0, 0, 1, 12) - AIC:1956.2614616844573
SARIMA(0, 1, 0)x(0, 0, 2, 12) - AIC:1723.153364023447
SARIMA(0, 1, 0)x(0, 0, 3, 12) - AIC:4047.750993647416
SARIMA(0, 1, 0)x(1, 0, 0, 12) - AIC:1837.4366022456677
SARIMA(0, 1, 0)x(1, 0, 1, 12) - AIC:1806.990530138882
SARIMA(0, 1, 0)x(1, 0, 2, 12) - AIC:1633.2108735791837
SARIMA(0, 1, 0)x(1, 0, 3, 12) - AIC:4198.7747798418895
SARIMA(0, 1, 0)x(2, 0, 0, 12) - AIC:1648.3776153470858
SARIMA(0, 1, 0)x(2, 0, 1, 12) - AIC:1647.2054158613616
SARIMA(0, 1, 0)x(2, 0, 2, 12) - AIC:1630.9898053920804
SARIMA(0, 1, 0)x(2, 0, 3, 12) - AIC:3534.298287308879
SARIMA(0, 1, 0)x(3, 0, 0, 12) - AIC:1467.4574095308406
SARIMA(0, 1, 0)x(3, 0, 1, 12) - AIC:1469.1871052625634
SARIMA(0, 1, 0)x(3, 0, 2, 12) - AIC:1471.0594530064295
SARIMA(0, 1, 0)x(3, 0, 3, 12) - AIC:1660.0206716336645
SARIMA(0, 1, 1)x(0, 0, 0, 12) - AIC:2230.162907850583
SARIMA(0, 1, 1)x(0, 0, 1, 12) - AIC:1923.7688649566603
SARIMA(0, 1, 1)x(0, 0, 2, 12) - AIC:1692.7089572783755
SARIMA(0, 1, 1)x(0, 0, 3, 12) - AIC:3276.943173168113
SARIMA(0, 1, 1)x(1, 0, 0, 12) - AIC:1797.179588183827
SARIMA(0, 1, 1)x(1, 0, 1, 12) - AIC:1738.0903193744662
SARIMA(0, 1, 1)x(1, 0, 2, 12) - AIC:1570.1509144550118
SARIMA(0, 1, 1)x(1, 0, 3, 12) - AIC:3799.5363478099143
SARIMA(0, 1, 1)x(2, 0, 0, 12) - AIC:1605.675195417545
SARIMA(0, 1, 1)x(2, 0, 1, 12) - AIC:1599.2245085437517
SARIMA(0, 1, 1)x(2, 0, 2, 12) - AIC:1570.4018823125118
SARIMA(0, 1, 1)x(2, 0, 3, 12) - AIC:3189.9386830433173
SARIMA(0, 1, 1)x(3, 0, 0, 12) - AIC:1428.4607679617193
SARIMA(0, 1, 1)x(3, 0, 1, 12) - AIC:1428.8727984071902
```

After the sort we found that Less the AIC we will take that model in this case (1,1,1,) (0,0,3,12) has the lowest AIC.

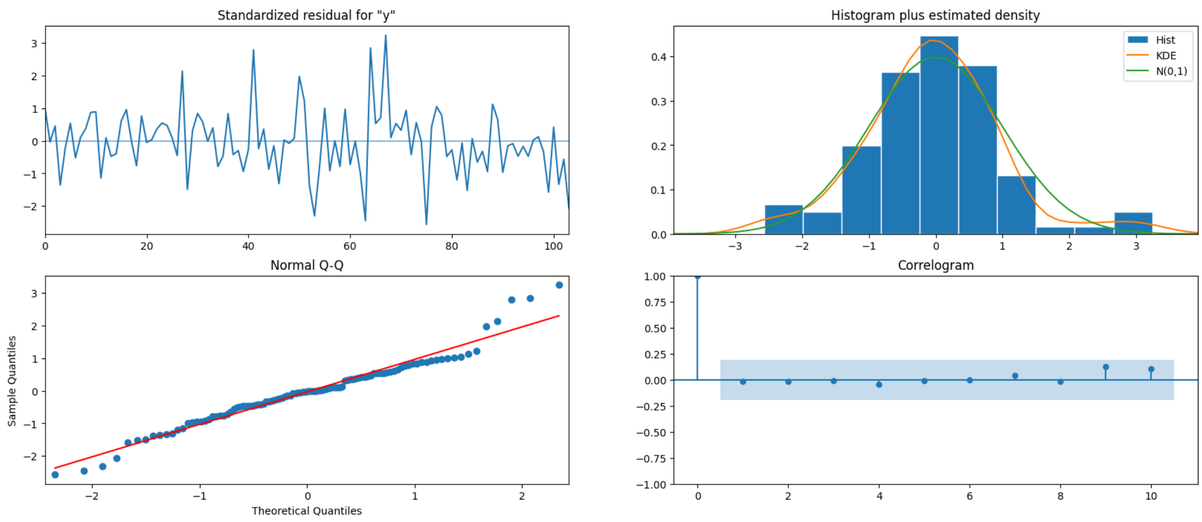
	param	seasonal	AIC
83	(1, 1, 1)	(0, 0, 3, 12)	12.000000
163	(2, 1, 2)	(0, 0, 3, 12)	16.000000
115	(1, 1, 3)	(0, 0, 3, 12)	16.000000
179	(2, 1, 3)	(0, 0, 3, 12)	967.881216
252	(3, 1, 3)	(3, 0, 0, 12)	1387.497014

The summary report for the ARIMA Model with values (1,1,1,) (0,0,3,12) model.

```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  132
Model:             SARIMAX(1, 1, 2)x(1, 0, 2, 12)   Log Likelihood                -770.792
Date:                            Sun, 25 Feb 2024   AIC                           1555.584
Time:                                    16:30:41   BIC                           1574.095
Sample:                                         0   HQIC                          1563.083
                                            - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.6282      0.255     -2.464      0.014      -1.128      -0.128
ma.L1         -0.1040      0.225     -0.463      0.644      -0.545       0.337
ma.L2         -0.7276      0.154     -4.736      0.000      -1.029      -0.427
ar.S.L12       1.0439      0.014     72.839      0.000       1.016       1.072
ma.S.L12      -0.5550      0.098     -5.663      0.000      -0.747      -0.363
ma.S.L24      -0.1354      0.120     -1.133      0.257      -0.370       0.099
sigma2      1.506e+05   2.03e+04      7.401      0.000    1.11e+05     1.9e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.04   Jarque-Bera (JB):                11.72
Prob(Q):                              0.84   Prob(JB):                         0.00
Heteroskedasticity (H):               1.47   Skew:                             0.36
Prob(H) (two-sided):                  0.26   Kurtosis:                         4.48
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

Graphs for the residual to determine if any further information can be extracted or all the usable information has already been extracted .



y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	1327.325736	388.341704	566.189983	2088.461489
1	1315.072585	402.007683	527.152005	2102.993164
2	1621.527819	402.001296	833.619758	2409.435880
3	1598.823447	407.240881	800.645987	2397.000907
4	1392.635677	407.971310	593.026602	2192.244752

- **RMSE for SARIMA:**

```
528.6593092642535
```

7. Build a table (create a data frame) with all the models built along with their corresponding parameters and the respective RMSE values on the test data.

As we can see that the model that has the lowest RMSE is Exponential Smoothing with 0.4 as Alpha , 0.1 as beta and 0.3 as Gamma with 378.62 is the best.

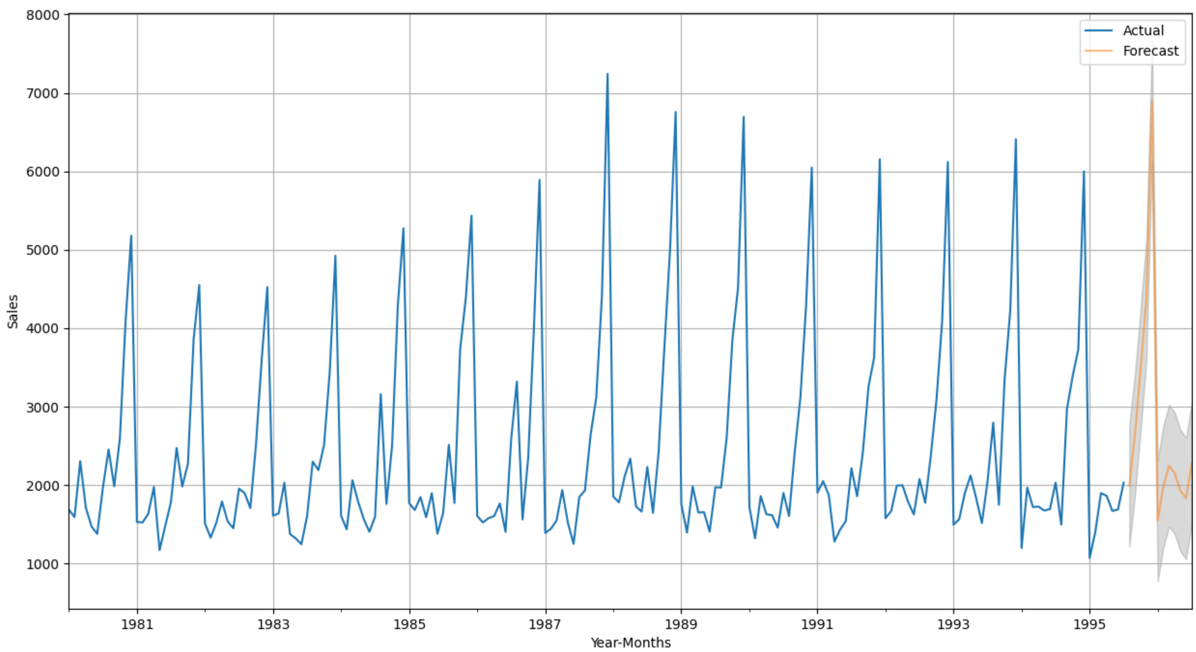
	RMSE
Alpha=0.4,Beta=0.1,Gamma=0.3:TES	378.626241
(1,1,1),(2,0,3,12),Auto_SARIMA	528.659309
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
ARIMA(2,0,1)	1269.345658
SimpleAverageModel	1275.081804
6pointTrailingMovingAverage	1283.927428
ARIMA(2,1,2)	1299.979749
ARIMA(3,1,3)	1319.936734
Alpha=0.1,SES	1338.004623
9pointTrailingMovingAverage	1346.278315
RegressionOnTime	1389.135175
NaiveModel	3864.279352
Alpha=0.1,Beta=0.1:DES	5291.879833

8. Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.

We can see that the optimum model with lowest RMSE is exponential smoothing so this model will be ideal for making predictions. Considering Exponential smoothing model ideal we will make prediction as below:

	Sales_Predictions
1995-08-01	1988.782193
1995-09-01	2652.762887
1995-10-01	3483.872246
1995-11-01	4354.989747
1995-12-01	6900.103171
1996-01-01	1546.800546
1996-02-01	1981.361768
1996-03-01	2245.459724
1996-04-01	2151.066942
1996-05-01	1929.355815
1996-06-01	1830.619260
1996-07-01	2272.156151

The Sales prediction of Sparkling wine graph with confidence level as shown below:



9. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.

- The sales for Sparkling wine for the company are predicted to be at least the same as last year, if not more, with peak sales for next year potentially higher than this year.
- Sparkling wine has been a consistently popular wine among customers with only a very marginal decline in sales, despite reaching its peak popularity in the late 1980s.
- Combining promotions where Sparkling wine is paired with a less popular wine such as "Rose wine" under a special offer may encourage customers to try the underperforming wine, which could potentially boost its sales and benefit the company
- Seasonality has a significant impact on the sales of Sparkling wine, with sales being slow in the first half of the year and picking up from August to December.
- It is recommended for the company to run campaigns in the first half of the year when sales are slow, particularly in the months of March to July.